Why Do We Need to Employ Bayesian Statistics and How Can We Employ it in Studies of Moral Education?: With Practical Guidelines to Use JASP for Educators and Researchers

Hyemin Han

University of Alabama

Joonsuk Park

The Ohio State University

Stephen J. Thoma

University of Alabama

Author Note

Hyemin Han (1-205-348-0746, hyemin.han@ua.edu) and Stephen J. Thoma (1-205-348-8146, sthoma@ua.edu), Educational Psychology Program, University of Alabama, Box 870231, Tuscaloosa AL 35487, USA.

Joonsuk Park (1-614-292-8185, park.1952@buckeyemail.osu.edu), Department of Psychology, The Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus OH 43210, USA.

Correspondence concerning this article should be addressed to Hyemin Han, Educational Psychology Program, University of Alabama, Tuscaloosa, AL 3547. This research was conducted at the University of Alabama.

Contact: hyemin.han@ua.edu

Why Do We Need to Employ Bayesian Statistics and How Can We Employ it in Studies of Moral Education?: With Practical Guidelines to Use JASP for Educators and Researchers

Abstract

In this paper, we discuss the benefits of and how to utilize Bayesian statistics in studies of moral education. To demonstrate concrete examples of the applications of Bayesian statistics to studies of moral education, we reanalyzed two datasets previously collected: one small dataset collected from a moral educational intervention experiment, and one big dataset from a large-scale Defining Issues Test-2 survey. Results suggest that Bayesian analysis of datasets collected from moral educational studies can provide additional useful statistical information, particularly that associated with the strength of evidence supporting alternative hypotheses, which has not been provided by the classical frequentist approach focusing on *P*-values. Finally, we introduce several practical guidelines pertaining to how to utilize Bayesian statistics, including the utilization of newly developed free statistical software, Jeffrey's Amazing Statistics Program (JASP), and thresholding based on Bayes Factors, to scholars in the field of moral education.

Keywords: Bayesian statistics, Bayes Factor, P-values, Statistical analysis, JASP

Introduction

The Journal of Moral Education (JME) aims to introduce cutting-edge interdisciplinary studies contributing to the improvement of moral education (Kristjánsson, 2017). In order to achieve the aforementioned purpose, it is important to facilitate vigorous interactions between diverse fields contributing to moral education, including but not limited to moral philosophy, moral psychology, and moral education (Han, 2014; Jeong & Han, 2013). Particularly, developing effective educational programs and activities for moral education promoting moral development and positive youth development can be realized by constructing morally-valid

conceptual foundations based on moral philosophy and designing concrete activity components based on scientific evidence collected in the field of moral psychology (Han, 2014).

Hence, we, researchers and educators in moral education, should pay keen attention to how to gather and test evidence supporting empirical aspects of moral education. Because educational activities, even those conducted during a very short period, can alter the developmental trajectories of students, such as the longitudinal changes in their academic achievement and well-being, significantly for the long term (Yeager & Walton, 2011), we need to be careful while trying to apply findings from empirical research to educational practice. Thus, we need to employ reliable and valid methodologies to analyze data for moral education (Han, Lee, & Soylu, 2016); it would be particularly important in the field of moral education, because moral education can significantly influence students' behavior in social and civic contexts and well-being (Althof & Berkowitz, 2006; Han, 2015; Kristjánsson, 2013).

Although recent debates about the frequentist perspective in the field of quantitative methods have intensified concerns regarding how to collect and test data properly (Benjamin et al., 2018), the majority of studies in the fields related to moral education have tend to use such a perspective. We have been used to employing the methodology of frequentist, such as *P*-values, in empirical studies of moral education. For instance, when we searched for articles published in the JME using a keyword, "Bayesian," which has been considered to be able to address limitations of the frequentist approach, only four items were found[i]. In fact, none of them actually used Bayesian inference in addition to or in place of classical frequentist inference; instead, three of them utilized Bayesian Information Criteria for model selection (Aho, Derryberry, & Peterson, 2014), and one merely cited another previous study using Bayesian approach. Furthermore, we searched for more peer-reviewed articles related to this topic from

the PsycInfo, the database of psychological articles organized by the American Psychological Association (APA). We used two keyword sets, "Bayesian" AND "Moral Education," and "Bayesian" AND "Moral Development," to search for previous articles relevant to the topics of the JME. When we entered the first keyword set, "Bayesian" AND "Moral Education," we found only one article published so far [ii].When we used the second keyword set, "Bayesian" AND "Moral Development," the PsychInfo returned four peer-reviewed articles [iii].

Among the aforementioned five articles extracted from the PsycInfo, two articles authored by Walker, Gustafson and Hennig (2001), and Walker, Gustafson, and Frimer (2007) provide useful insights about how Bayesian approach can contribute to our field, moral development in particular. Walker et al. (2001) employed Bayesian analysis to examine how developmental data was classified, because classical frequentist methods were not appropriate for this purpose as classifying developmental data is based on probabilistic assumptions. Their study demonstrated that Bayesian analysis can help us better analyze probabilistic developmental classifications that could not be feasibly done with frequentist methods. Moreover, Walker et al.'s (2007) review article described benefits of Bayesian analysis in developmental psychological studies. They suggested that first, the interpretation of results from Bayesian analysis is more intuitive than that of p-values; second, it is easier to address missing data and measurement error issues with Bayesian techniques; third, Bayesian techniques can be feasibly employed to analyze hierarchical models associated with unobserved or latent structures (Walker et al., 2007).

Although these previous studies were able to demonstrate potential benefits of Bayesian analysis in studies of moral education and moral development with a concrete example (Walker et al., 2001) and give us a primer on it values (Walker et al., 2007), they could not provide

educators and researchers, who are unlikely to have sufficient statistical expertise, with practical guidelines about how to implement Bayesian analysis. For example, Kondo's (1990) previous Bayesian modeling of moral behavior used MPlus (Muthén & Muthén, 2011). Walker et al. (2007) suggested researchers use WinBugs (MRC Biostatistics Unit, 2018). Although the aforementioned statistical tools provide users with very powerful functionalities, users should have expertise in statistics and computer programs as they require text-based coding. Instead, many moral educators and researchers might be familiar with simple tools featured with graphical user interface, such as SPSS (IBM, 2018). In addition, the aforementioned articles did not show how to perform basic statistical tests, such as *t*-tests, ANOVA, and correlation analysis, that many educators and researchers are interested in with Bayesian techniques.

Thus, we intend to consider why we should employ Bayesian techniques in studies of moral education, and how to feasibly utilize such techniques. In fact, we, researchers and educators in the field of moral education have rely heavily on the frequentist approach, which is currently being criticized by quantitative methodologists, but have not paid much attention to alternative statistical methods, particularly Bayesian methods, that can address the limitations of the approach. In addition, although the aforementioned previous studies found from the PsycInfo have employed Bayesian statistics and contributed to the methodological improvements in the field, they could not provide concrete and feasible guidelines that can be easily understood by educators and scholars who do not have expertise in statistics. Hence, we overview Bayesian analysis and its benefits, examine how such analysis works with concrete datasets, and propose feasible practical guidelines for educators and researchers.

A Brief Introduction to Bayesian Analysis

To understand what Bayesian analysis is, it would be helpful to contrast it to the mainstream, traditional statistical approach: frequentist statistics. The fundamental difference between frequentist and Bayesian schools of statistics comes from the different ways they view quantities of interest or parameters. Frequentists view parameters such as population mean or variance as 'fixed but unknown constants' (Neyman, 1977); the notion of randomness does not apply to the parameters themselves. In other words, parameters cannot be considered random. For example, when a researcher constructs a 95% confidence interval about the population mean, she is not saying, in principle, that the true mean is contained in the specific interval she constructed with a probability of 95%. This is prohibited in the frequentist school because we cannot give a probabilistic interpretation to a single event, i.e., a single confidence interval. Instead, '95%' is a quality of the confidence-interval-construction-procedure itself in that, when constructed repeatedly under repeated experimentation, 95% of such intervals constructed would *contain* the true mean. Since we cannot make direct probabilistic statements about the parameters, it is difficult to express our uncertainty or degree of belief (or doubt) directly about them.

On the other hand, Bayesians allow for expressing randomness about parameters themselves. That is, they regard parameters as random quantities or variables, not as fixed. It is possible to assume the probability distribution of a parameter of interest such as population means or differences among them. This allows researchers to express uncertainty about parameters in the form of probability distribution. The goal of Bayesian inferences is then to 'update' the distribution in light of the data at hand. Bayes' theorem is the universal mathematical tool for that purpose, which is formulated as

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where θ denotes parameter(s). Description of each term is in order. For simplicity, we will assume that the parameter space is discrete for a moment. First, P(θ) is called the 'prior probability' and represents the probability that the parameter value is equal to θ *before* observing the data. And P(θ|D) is called 'posterior probability' which is the probability that the parameter value is θ *after* observing the data. In fact, the above formula can be understood as showing how prior is updated to be the posterior: posterior is prior times P(D|θ)/P(D) where P(D|θ) is called *likelihood* (the same concept as in frequentist statistics) and P(D) the *marginal* probability. Through applying Bayes' theorem using the data at hand, which is known as *Bayesian updating*, one can update the probability that the parameter value is equal to each possible value of θ, yielding the *posterior distribution* of the parameter. The posterior distribution represents the updated probability (or degree of belief) about the parameter of interest after observing the data. In case of continuous parameters, probability masses are replaced by probability densities, but the general framework is not much different from the discrete case.

Here is a basic example of how Bayesian inference works. Say we are interested in obtaining an interval estimate about a normal mean where we know the population variance. Given independent and identically distributed data, $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, where σ^2 is known, the traditional frequentist approach would yield a 95% confidence interval about μ as $[\bar{X} - z_{.975}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{.975}\frac{\sigma}{\sqrt{n}}]$ where $\bar{X} = \frac{1}{n}\sum X_i$ is the sample mean and $z_{.975}$ is the 97.5th percentile of the standard normal distribution. Note that this interval does not make a probabilistic statement directly about μ. Indeed, introductory statistics textbooks warn against concluding that μ is in the constructed interval with a 95% probability.

How does a Bayesian approach differ from the frequentist one? First, Bayesian updating should be done. Here we will take the prior distribution for μ to be $N(m, s^2)$. This reflects our prior belief or expectation that μ is close to m, and the degree of uncertainty is quantified by s^2. These values should be chosen by the researcher in advance. There are many ways of doing this which we will not discuss in further detail here. Using the prior distribution and the data as described before, it can be shown that, by making use of Bayes' theorem, the posterior distribution of μ is given by $N(m', (s)^2)$ where $m' = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+s^2} m + \frac{s^2}{\frac{\sigma^2}{n}+s^2} \bar{X}$ and $(s')^2 = \left(\frac{n}{\sigma^2} + \frac{1}{s^2}\right)^{-1}$ where m' is called the posterior mean and $(s')^2$ the posterior variance (Bolstad & Curran, 2016). These quantities reflect our modified belief about μ after observing the data. Note that the posterior variance, $(s')^2 = \left(\frac{n}{\sigma^2} + \frac{1}{s^2}\right)^{-1}$, is smaller than the usual frequentist standard error, $\frac{\sigma^2}{n}$, which is due to the fact that we exploited the prior information about μ. One can use the posterior distribution to construct an interval estimate for μ, $[m' - z_{.975}\frac{s'}{\sqrt{n}}, m' + z_{.975}\frac{s'}{\sqrt{n}}]$ as in case of the frequentist approach. Bayesians typically refer to this type of interval estimates as *credible interval*s (Edwards, Lindman, & Savage, 1963). The interpretation of credible intervals differ from that of confidence intervals because we are allowed to make direct probabilistic statements about μ. For example, it is legitimate to say that the probability that μ lies in the constructed credible interval is equal to 0.95 because Bayesian interpretation of probability does not require the notion of repetition as frequentists do, and parameters are regarded as random variables.

Bayesian analysis is believed to have the potential to address many problems associated with the (mis)uses of *P*-values in research (Wagenmakers, 2007). Discussions on this topic

abound, but we would like to briefly summarize the reason why Bayesian analysis is in the limelight recently. To begin with, here are specific problems of using *P*-values. First, *P*-values do not convey the information researchers would like to learn about, namely, the probability that a hypothesis is true given data, P(H|D) (Cohen, 1994; Wagenmakers, 2007). What *P*-values actually mean is, however, the inverse of that probability. That is, according to the definition of *P*-values, they are defined as the probability of observing data as extreme or more extreme than the one observed, P(D|H). This is not the same as P(H|D), as Bayes' theorem tells us. Because the use of Bayes' theorem itself does not hinge on whether one is frequentist or Bayesian, this criticism is not limited to the Bayesians' perspectives. In addition, regarding this problem, misunderstandings about *P*-values are widespread (Gigerenzer, 2004), which exacerbates the problem.

Second, *P*-values can be easily 'hacked' (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Simmons, Nelson, & Simonsohn, 2011). That is, researchers can engage in questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012) to obtain 'significant' P-values that are less than .05, which is accepted widely by convention (Nelson, Simmons, & Simonsohn, 2018). For example, if they observe a *P*-value which is slightly greater than .05, they could simply collect more data and conduct test again, and do this repeatedly until they reach $P < .05$. This practice is known as 'optional stopping' in the literature, and is thought to inflate Type 1 error (Rouder, 2014). *P*-hacking is thought to impact the empirical distribution of *P*-values observed across the entire field, and there is evidence implying that it is actually happening (Simmons et al., 2011).

Then, how does Bayesian analysis address those problems? First, Bayesian analysis provides researchers with a way to compute the quantity they are ultimately interested in, P(H|D),

the probability that some hypothesis is true given data at hand. Frequentists cannot define or compute this quantity since they are not allowed to assign probabilities of being true to hypotheses. Frequentists can assign a probability only to events from repeatable experiments, which is not the case for scientific hypotheses. Bayesians can do this, however, because a probability can be defined as 'degree of belief' about some event, which makes possible to assign a probability to a single event, which in the context of research is 'the null/alternative hypothesis is true.' So, they can talk about P(H) or P(H|D), the probabilities that a hypothesis is true before or after observing the data.

Second, it is not yet clear how to 'hack' a Bayes Factor. At least, it comes with some forms of protection against *P*-hacking. First, researchers cannot simply increase sample sizes to obtain 'significant' results. In Bayesian hypothesis testing, interestingly, for a fixed value of *P*-value, the support for a null *increases*, not decreases, as the sample size increases. The following table is taken from Berger and Sellke (1987).

<Place Table 1 about here>

We can see that, within a single row, $P(H_0|D)$ increases as sample size increases. For example, when $P = .05$, which is the gold standard researchers use in most behavioral research, one can see that $P(H_0|D) = .35$ when $n = 1$, but it becomes .82 when $n = 1{,}000$. That is, the null is more likely to be true when $n = 1{,}000$ than $n = 1$. This seems to run against the common sense that, for a fixed effect size, statistical results become more 'significant' as *n* grows. It turns out that Bayesian hypothesis testing has a built-in inner structure that penalizes more complex hypothesis (models), and in this case the more 'complex' one was the alternative hypothesis because it says that the true parameter value can be anything but zero whereas the null that it must be exactly equal to zero. As a result, the same *P*-value obtained from a larger sample would

provide more support for the null than one from a smaller sample, and attempts to 'hack' Bayes Factors is not likely to be successful.

In Bayesian analysis, Bayes Factors (BFs) are a popular tool to quantify evidence of the null against the alternative, or vice versa. A Bayes Factor is defined as follows:

$$BF_{10} = \frac{P(D|H_1)}{P(D|H_0)} = \frac{P(H_1|D)/P(H_1)}{P(H_0|D)/P(H_0)}$$

where $P(D|H_0)$ and $P(D|H_1)$ are probabilities of observing data under the null, and the alternative, respectively. The expression above also tells us that a BF is also the ratio of the posterior odds, $P(D|H_1)/P(D|H_0)$, to the prior odds, $P(H_1)/P(H_0)$. That is, a BF is the factor by which the prior odds is multiplied to yield the posterior odds. Since prior odds and posterior odds are the ratio of beliefs about the hypotheses before and after observing the data, BFs can be thought of as quantifying how much one should adjust the prior belief give the data. Guidelines for interpreting BFs exist (Jeffreys, 1961; Kass & Raftery, 1995).

Thanks to the advantages of conducting Bayesian analyses instead of traditional, frequentist analyses, discussions about them are burgeoning now. What is lacking from the academic discourses are how to actually conduct them in practice. As pointed out, only few articles mentioned and utilized Bayesian analysis. Although some of them, particularly Walker et al.'s (2001) study showed how Bayesian techniques can contribute to probabilistic developmental classifications with a concrete example, the lack of practical guidelines for end users in the previously published articles can be problematic to help the end users employ Bayesian analysis in their studies of moral education. This has to be addressed as well.

Present Study

We aim to introduce the perspective and methodology of Bayesian statistics to readers of the JME, researchers and educators in moral education, with concrete examples and practical

guidelines. Particularly, we intend to introduce a new tool implementing Bayesian inference, Jeffrey's Amazing Statistics Program (JASP)[iv], which has a feasible user interface (Love et al., 2017), can be utilized conveniently for studies of empirical psychology (Marsman & Wagenmakers, 2017; Wagenmakers, Love, et al., 2017). Also, we plan to focus on basic statistical methods that have been frequently utilized by moral educators and researchers (e.g., *t*-test and ANOVA, correlation analysis), but have not been well introduced in the previous published articles pertaining to Bayesian analysis and moral educational studies. First, we will compare outcomes from the classical (frequentist) and Bayesian analyses of data collected from the previous psychological studies of moral education. Second, we will discuss some Bayesian benefits and practical guidelines about how to perform Bayesian inference with JASP for studies of moral education with screenshots demonstrating how to set JASP options to perform various types of Bayesian analyses.

Method

Materials

In order to compare outcomes from statistical analyses based on the frequentist and Bayesian perspective, we reanalyzed two datasets pertaining to moral development and moral education collected by two previous studies.

The first reanalyzed dataset contained data collected from the previous studies that compared motivational effects of the stories of peer moral exemplars and historic figures among 111 Korean eighth graders. The original findings, which resulted from classical statistical analyses, were reported in Study 2 in Han, Kim, Jeong and Cohen (2017). In this study, the students participated in two different types of moral educational intervention activities for eight weeks. On the one hand, the peer exemplar group was asked to discuss and praise moral virtues

and behaviors of their peer moral exemplars, such as family members and friends. On the other hand, the historic figure group discussed and praises moral virtue and actions of historic moral exemplars, such as Mother Teresa and Martin Luther King, Jr. The intervention session was conducted once a week for an hour. Before the beginning of the intervention period, Han et al. (2017) surveyed the students' initial engagement in voluntary service activities. The same variable was surveyed once again twelve weeks after the pre-test survey period. Han et al. (2017) reported that the post-test service engagement in the peer exemplar group was greater than that in the historic figure group after controlling for the pre-test service engagement.

The second dataset was collected from national norming sample of 32,229 college students who completed the Defining Issues Test-2 during the 2010-2015 timeframe (DIT-2; Rest, Narvaez, Thoma, & Bebeau, 1999; Center for the Study of Ethical Development). The DIT-2 consists of 5 dilemmas; each followed by 12 items. Participants are first asked to take the role of the protagonist in the story and decide what he/she ought to do, and are then asked to rate and rank the items in terms of their importance in interpreting the moral dilemma. Traditionally, the summary score of the DIT has been the "P" score, which is interpreted as the relative importance given to post-conventional (i.e., Kohlberg's Stages 5 and 6) moral considerations. The newer N2-score used in this study is an improvement over the P-score as an overall estimate of moral judgment development (Thoma, 2006). More recently, the construct measured by the DIT has been reinterpreted (Rest, Narvaez, Bebeau, & Thoma, 1999a, 1999b). Based upon large-sample analyses, it appears that the DIT measures three developmentally ordered schema: personal interest (incorporating aspects of Kohlberg's stages two and three), maintaining norms (closely aligned with Kohlberg's stage 4) and post-conventional schema (the traditional P score

mentioned above). The validity and reliability of the DIT is fully discussed in Rest, Narvaez, Bebeau, et al. (1999b; see also Thoma, Bebeau, Dong, Liu, & Jiang, 2011).

Procedures

We reanalyzed the two datasets collected by the previous moral educational studies with JASP. First, we conducted both classical and Bayesian *t*-tests, ANCOVA, and correlation analysis with the moral exemplar intervention data. In the case of the *t*-tests, we compared the changes in voluntary service engagement (post-test – pre-test) between the peer moral exemplar and historic figure groups. For readers' practical guidelines, we demonstrate how to set JASP options to perform Bayesian t-test and ANCOVA in screenshots (see Figure 1). the case of the ANCOVA, we set the change in voluntary service engagement as a dependent variable, the group assignment as a fixed factor, and the pre-test engagement as a covariate. While performing the Bayesian ANCOVA, we examined which model was the best model among possible models by comparing BFs. We used default values pre-set by JASP for priors and other parameters. For these analyses, we intended to see whether the peer exemplar group better increased service engagement compared with the historic moral exemplar group.

Second, we performed both classical and Bayesian ANCOVA and correlation analyses with the DIT-2 dataset. For the ANCOVA, we set the N2-score as a dependent variable, and sex and grade level as fixed factors. In this process, the interaction effect between sex and grade level was also examined. For the correlation analyses, we examined correlation among the P-score (to explore association with N2-score), N2-score, age, and grade level. Again, to provide readers with practical guidelines, we present how to use JASP to perform Bayesian correlation analysis and ANCOVA in screenshots (see Figure 2). Similar to the Bayesian analyses of the moral educational intervention data, we used the default setting provided by JASP. We also

examined which model was the best model by comparing calculated BFs. By performing these analyses, we would like to test whether the aforementioned demographical factors, sex, age, and grade level, were associated with P- and N2-scores. Previous research has shown that there are significant associations among P- and N2-scores, age and grade level in college by conducting classical ANOVA and correlation analysis (Rest, Narvaez, Thoma, et al., 1999). In addition, there have been debates about whether any gender biases, particularly those that are likely to penalize women, are embedded in the DIT and Kohlbergian perspective (Gilligan, 1982; Thoma, 1986). However, they have not utilized Bayesian techniques, so we intended to test the aforementioned matters with Bayesian techniques to examine how they performed compared with frequentist techniques.

In the cases of classical analyses, we used *P*-values as indicators for significance (i.e., $p < .05$, $p < .01$, and $p < .001$). In the cases of Bayesian analyses, we examined BFs, more specifically the natural logarithm values of BF, logBF, as indicators for the strength of evidence supporting H_1 instead of H_0. Following the guidelines recommended by Kass and Raftery (1995), we used 2logBF = 2 for the threshold of positive evidence, 2logBF = 6 for the threshold of strong evidence, and 2logBF = 10 for the threshold of very strong evidence.

Results

First, we reanalyzed the moral educational intervention dataset with JASP. When we performed a classical *t*-test, the result reported that the increase in voluntary service activity engagement was greater in the peer exemplar group than that in the historic figure group, $t(103) = -2.66$, $p < .01$, $D = -.52$. However, unlike the result of the classical *t*-test with $p < .01$, the result from the Bayesian *t*-test indicated that there was only weakly positive evidence supporting H_1 instead of H_0, $BF_{10} = 4.61$, 2logBF = 3.06 (see Figure 1 for the prior and posterior distribution).

Moreover, the result from classical ANCOVA indicated that both the group assignment, F (1, 102) = 8.58, $p < .01$, and pre-test service engagement, F (1, 102) = 2.15, $p < .01$, influenced the change in service engagement. Bayesian model comparison reported that the model with both main effects of group assignment and pre-test engagement was the best model (see Table 2). The result indicated that there was strong evidence to select the best model instead of the null model. However, unlike the result from classical ANCOVA, the result from Bayesian ANCOVA with the best model showed that evidence supporting the presence of the effect of group assignment was positive but not strong, while that supporting the presence of the effect of pre-test engagement was stronger, (see Table 3).

<Place Figure 1, and Tables 2 and 3 about here>

Second, we performed classical and Bayesian ANCOVA and correlational analysis of the DIT-2 dataset. Although both main effects of sex, F (1, 32,221) = 269.06, $p < .001$, and grade level were significant, F (3, 32,221) = 50.13, $p < .001$, the interaction effect between these two main effects was marginal according to the result from classical ANCOVA, F (3, 32,221) = 2.15, $p = .09$. Bayesian model comparison reported that the model only with both main effects was the best model; instead, the model including the interaction effect was considered less favorable than the best model (see Table 5). The result of Bayesian ANCOVA was consistent with the results of the classical ANCOVA and Bayesian model comparison. There was very strong evidence supporting the inclusion of the main effects of sex and grade level to the analysis model; however, findings suggested we exclude the interaction effect between the two main effect from the model (see Table 6). The results of classical and Bayesian correlation analyses are presented in Table 7. All correlation coefficients were found to be significant from the classical correlation analysis. Similarly, all calculated 2logBF values exceeded the threshold of very strong evidence

(2logBF = 10). However, we were able to discover that indicated strengths of evidence were very diverse across different associations (e.g., 2logBF = 13.94 in the case of N2-score and age vs. 2logBF = ∞ in the case of P-score and N2-score). The results suggested that first, both grade level and sex significantly predicted a P-score; grade level and being a woman was positively associated with the score. Second, grade level was positively associated with both P- and N2-scores, but age was negatively associated with the scores; however, the strength of evidence supporting the presence of association was weaker in the case of age compared with grade level.

<Place Tables 5, 6, and 7 about here>

Discussions

In the present study, we compared outcomes from classical and Bayesian statistical analyses of datasets originally collected by two previous studies of moral education with JASP. The analyses of the moral educational intervention dataset demonstrated that although findings from classical analyses indicated the significance of both factors, group assignment and pre-test engagement, at $p < .05$ or even at $p < .01$, Bayesian analyses reported weakly positive evidence supporting H_1 instead of H_0. Findings from the analyses of the DIT-2 dataset corroborates such a point. The results from the correlation analyses showed that although classical analysis indicated all associations as very significant ($p < .001$), the result of Bayesian analysis can show us how the strength of evidence supporting each association was different across different associations with BFs (from 2logBF = 13.94 to ∞). Moreover, the Bayesian ANCOVA provided us with more information regarding which model should be employed by comparing BF values from different models.

In the case of the reanalysis of the moral educational intervention dataset, we found a divergence between the outcomes from classical and Bayesian analysis. In this situation, practically, both divergent outcomes, including the unfavorable outcome from Bayesian analysis, might need to be reported when Bayesian techniques are applied. Researchers may consider reporting something like the following: "Although the result from classical inference indicated that effect of the factor was significant ($p < .05$), Bayesian inference showed that the supporting evidence is rather weak ($2 \leq 2\log BF < 6$) (see Kass and Raftery (1995) for criteria). Thus, we need to interpret the effect of the factor with a caution." Of course, some may argue that we can still criticize the weak outcome from the reanalysis based on the contemporary frequentist perspective. The reanalyzed study was a study to design and test interventions in a classroom scale, so it recruited a small number of participants. Thus, such a situation might lead to a relatively large effect size with a small sample size, and we might not be able to accept the original conclusion ($p < .05$ and $p < .01$) very confidently (Begg, 1994).

Although we may be able to criticize the original findings from the contemporary frequentist perspective, Bayesian perspective can make their interpretation more straightforwardly; such a point will provide practical benefits to educators and researchers. Even if we can address issues related to *p*-values by reporting effect sizes and other additional statistical indicators, because a *p*-value has been regarded as an indicator that can be interpreted very simply ($p < .05$), non-experts might still be attracted by such a point (Cohen, 1994). We will need to interpret multiple indicators (e.g., effect sizes, sample sizes, etc.), which might seem to be less straightforward to interpret among non-experts compared with *p*-values, to make a better judgment from the frequentist perspective. In addition, we still cannot have any direct information about whether the collected data supports our hypotheses with the reported

frequentist indicators; such indicators inform us whether it is possible to reject null hypotheses (Kruschke & Liddell, 2018). If we utilize Bayesian techniques, we simple need to interpret one indicator, a Bayes factor, based on the suggested threshold values (e.g., Kass & Raftery, 1995). Also, it allows us to directly evaluate whether evidence supports our hypothesis instead of null hypotheses.

In the case of the second reanalysis, the reanalysis of DIT-2 data, we found significant associations among P- and N2-scores, and demographic factors, sex, age, and grade level. Although the results from frequentist analyses unequivocally reported very significant *p*-values, $p < .001$, Bayesian analyses reported varying degrees of the strength of evidence supporting different associations as shown by Bayes factors, 2logBF, ranging from 14 to higher than 100. Although we could not find any explicit contradicts between results from frequentist and Bayesian analyses unlike the case of the first reanalysis, Bayesian analysis was better in differentiating the strength of the significant or supportiveness of evidence compared with frequentist analysis. Such a point will provide educators and researchers, future users of Bayesian techniques, with practical benefits related to the interpretation of outcomes.

However, there were several points that should be interpreted with caution from the second reanalysis. First, although P- and N2-scores, and age and grade level showed very strong association between each other, it is not surprising at all to see such results. Because the N2 index takes into account whether participants prefer post-conventional scheme, which is naturally related to the P index, the correlation between P- and N2-scores are supposed to be inflated. Second, we analyzed P- and N2-scores instead of raw scores, which has been analyzed by Walker et al. (2001). Walker et al. (2001) argued that Bayesian techniques allow us to directly analyze the stage-type raw scores instead of summarized P- and N2-scores, so they are more

powerful tools for longitudinal analysis that is involved the statistical tests of transitions. Although we also acknowledge such a point, the strength of Bayesian analysis in sophisticated longitudinal analysis, we decided to perform *t*-tests, ANOVA, and correlation analysis with JASP, because we intended to show how such statistical tests that have been most frequently utilized by educators and researchers in studies of moral education, who might not have sufficient statistical knowledge and computer programing skills, can be feasible performed with Bayesian techniques. We did such tasks by showing screenshots from JASP with brief explanations (screenshots) about how to operate the program, and concrete analysis results.

Given these, Bayesian methods will contribute to better analyses in the field of moral education, which should be founded firmly on scientific evidence. More specifically, we may consider several fundamental practical benefits of Bayesian methods studies of moral education.

First, as pointed out earlier, employing Bayesian analysis allows for more epistemologically straightforward interpretations of statistical evidence in the form of Bayes Factors. This situation is in sharp contrast to the case of frequentist hypothesis testing, which is largely dependent upon the use of *P*-values that are neither directly relevant to the researchers' purposes nor make theoretical sense. Bayes Factors, in contrast, lets us to unambiguously quantify how much the data supports the null or the alternative more than the other, which we argue would be highly helpful for researchers by preventing potential misunderstandings.

Second, Bayesian analysis lets us to incorporate prior information that are relevant to the research question into the statistical analysis. This can be done via constructing a prior distribution for quantities of interest. Bayesian analysis is the process of updating prior distribution, our beliefs about the parameters of interest or the null/alternative hypothesis before seeing the data, to posterior distribution, updated beliefs about them after observing the data.

Constructing prior distributions is not trivial because it requires careful examination of the research problem itself and available information at hand. However, when appropriately done, the process can facilitate better statistical inferences. For example, previous research could inform us about what the effect size of interest is likely to be. Researchers could use this information to construct a prior distribution about the effect size. Results from Bayesian analysis are compromises between information from the prior and the data; imposing strong priors on the parameters makes the result closer to the prior beliefs. Sometimes the restriction prior information provides is more obvious. If a researcher were to use a 7-point Likert scale to measure the dependent variable, she already knows that the mean difference between the treatment and control groups cannot exceed 7. This information could also be taken into account in the form of prior on the mean difference to prevent it from being greater than 7.

Third, as shown in the present paper, Bayes Factors have built-in protective mechanisms which favors simpler models (hypotheses). This, we argue, is a very important but largely unnoticed advantage of using Bayes Factors instead of *P*-values. As scientists, we would naturally prefer simpler explanations about the phenomena of interest than more complex ones. This is known as 'Occam's razor,' and is regarded as one of the gold standards in evaluating scientific theories (Myung, Balasubramanian, & Pitt, 2000). *P*-values fail in this regard miserably because they do not 'penalize' complex models such as H_1: $\theta \neq 0$ sufficiently, thereby leading researchers to choose overly complex models over simpler ones such as H_0: $\theta = 0$, with sufficiently large sample sizes. This is not a result that researchers would like to see. Bayes Factors, instead, are a tool for hypothesis testing which takes into account the relative complexities of competing hypotheses (models), at least implicitly (Myung & Pitt, 1997). As we

saw earlier, as sample size grows, the same *t*-value yields increasingly large support for the null, not the alternative. We have also seen that results that are believed to strongly support the alternative (e.g., the effect of the group assignment with $p = .004$ in Table 2) do not hold true when re-analyzed in the Bayesian fashion using JASP (e.g., the effect of the group assignment with 2logBF = 3.05, which only suggests presence of positive but not strong evidence supporting an alternative hypothesis, in Table 4). They are, we believe, not coincidences but the instances where the tendency of Bayes Factors to favor simpler models is in action.

A newly developed tool for Bayesian analyses, JASP, would be useful for researchers and educators in the field of moral education (Wagenmakers, Love, et al., 2017). Because it supports both classical and Bayesian analyses with a simple user interface, even end users who are not fluent in advanced statistics and programming, which have been used to be required for the implementation of Bayesian methods, will be able to perform Bayesian analyses and compare their outcomes with those from classical analyses conveniently. Moreover, thanks to the development of computational technology and resources so far, Bayesian inference that requires numerous iterative calculations and long time to perform in the past now can be performed with a personal computer easily (Gronau, Wagenmakers, Heck, & Matzke, 2017).

Here is one practical guideline for future research in moral education using JASP: reporting resultant BFs on top of or instead of *P*-values. From the perspective of frequentists, *P*-values only provide information pertaining to whether a null hypothesis about the extremity of an observed distribution can be rejected; they do not say anything about whether and how strongly evidence found from a specific study supports a hypothesis (Wagenmakers, Marsman, et al., 2017). Furthermore, as the current debates indicated, conventional *P*-value thresholds widely used in the field, particularly, $p < .05$, could only support very week or even could not support

the presence of positive evidence (Benjamin et al., 2018). Instead, BFs show us the strength of evidence; directly BF thresholds used in the field can also be considered as better thresholds to make practical decisions about accepting a specific hypothesis based on evidence (Kass & Raftery, 1995). Hence, reporting BFs will provide readers, particularly researchers and educators who are interested in developing evidence-based moral educational programs, potentially with more practical information regarding whether findings, suggestions, and arguments in a specific article are well supported by empirical evidence. Furthermore, we believe that employing aforementioned guidelines will significantly contribute to the improvement of statistical rigorousness of empirical studies of moral education as well as empirical articles in the JME.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. doi:10.1890/13-1452.1

Althof, W., & Berkowitz, M. W. (2006). Moral education and character education: their relationship and roles in citizenship education. *Journal of Moral Education*, *35*(4), 495–518. doi:10.1080/03057240601012204

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 399–409). New York, NY: Russell Sage Foundation.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. doi:10.1080/01621459.1987.10478397

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian Statistics, Third Edition*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781118593165

Cohen, J. (1994). The earth is round ($p < .05$): Rejoinder. *American Psychologist*, *50*(12), 1103–1103. doi:10.1037/0003-066X.50.12.1103

Derryberry, W. P., & Thoma, S. J. (2005). Functional differences: comparing moral judgement developmental phases of consolidation and transition. *Journal of Moral Education*. doi:10.1080/03057240500049372

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. doi:10.1037/h0044139

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. doi:10.1016/j.socec.2004.09.033

Gilligan, C. (1982). *In a Different Voice*. *Harvard University Press*. Cambridge, MA: Harvard University Press. doi:10.2307/2067520

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2017). *A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models using Warp-III Bridge Sampling*. Retrieved from https://psyarxiv.com/yxhfm/

Han, H. (2014). Analyzing theoretical frameworks of moral education through Lakatos's philosophy of science. *Journal of Moral Education*, *43*(1), 32–53. doi:10.1080/03057240.2014.893422

Han, H. (2015). Virtue ethics, positive psychology, and a new model of science and engineering ethics education. *Science and Engineering Ethics*, *21*(2), 441–460. doi:10.1007/s11948-014-9539-7

Han, H., Kim, J., Jeong, C., & Cohen, G. L. (2017). Attainable and Relevant Moral Exemplars Are More Effective than Extraordinary Exemplars in Promoting Voluntary Service Engagement. *Frontiers in Psychology*, *8*, 283. doi:10.3389/fpsyg.2017.00283

Han, H., Lee, K., & Soylu, F. (2016). Predicting long-term outcomes of educational interventions using the Evolutionary Causal Matrices and Markov Chain based on educational neuroscience. *Trends in Neuroscience and Education*, *5*(4), 157–165. doi:10.1016/j.tine.2016.11.003

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, *13*(3).

doi:10.1371/journal.pbio.1002106

Heng, M. A., Blau, I., Fulmer, G. W., Bi, X., & Pereira, A. (2017). Adolescents finding purpose: Comparing purpose and life satisfaction in the context of Singaporean and Israeli moral education. *Journal of Moral Education*, *46*(3), 308–322. doi:10.1080/03057240.2017.1345724

Hodge, K., & Gucciardi, D. F. (2015). Antisocial and Prosocial Behavior in Sport: The Role of Motivational Climate, Basic Psychological Needs, and Moral Disengagement. *Journal of Sport and Exercise Psychology*, *37*(3), 257–273. doi:10.1123/jsep.2014-0225

IBM. (2018). IBM SPSS Statistics. Retrieved from https://www.ibm.com/products/spss-statistics

Jeffreys, H. (1961). *Theory of Probability*. *Theory of Probability* (Vol. 2). Oxford, UK: Oxford University Press.

Jeong, C., & Han, H. (2013). Exploring the Relationship between Virtue Ethics and Moral Identity. *Ethics & Behavior*, *23*(1), 44–56.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.2307/2291091

Kondo, T. (1990). Some Notes on Rational Behavior, Normative Behavior, Moral Behavior, and Cooperation. *Journal of Conflict Resolution*, *34*(3), 495–530. doi:10.1177/0022002790034003006

Kristjánsson, K. (2013). *Virtues and vices in positive psychology: A philosophical critique*. New York, NY: Cambridge University Press.

Kristjánsson, K. (2017). Moral education today: Ascendancy and fragmentation. *Journal of Moral Education*, *46*(4), 339–346. doi:10.1080/03057240.2017.1370209

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-4

Lee, C.-T., Padilla-Walker, L. M., & Nelson, L. J. (2015). A person-centered approach to moral motivations during emerging adulthood: Are all forms of other-orientation adaptive? *Journal of Moral Education*, *44*(1), 51–63. doi:10.1080/03057240.2014.1002460

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., & Wagenmakers, E. J. (2017). JASP (Version 0.8.2). Amsterdam, The Netherlands: Jasp project. Retrieved from https://jasp-stats.org/

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*(5), 545–555. doi:10.1080/17405629.2016.1259614

McGrath, R. E., & Walker, D. I. (2016). Factor structure of character strengths in youth: Consistency across ages and measures. *Journal of Moral Education*, *45*(4), 400–418. doi:10.1080/03057240.2016.1213709

MRC Biostatistics Unit. (2018). *The BUGS Project*. London, UK. Retrieved from https://www.mrc-bsu.cam.ac.uk/software/bugs/

Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*(21), 11170–11175. doi:10.1073/pnas.170283897

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. doi:10.3758/BF03210778

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. doi:10.1146/annurev-psych-122216-011836

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131. doi:10.1007/BF00485695

Railton, P. (2017). Moral Learning: Conceptual foundations and normative relevance. *Cognition*, *167*, 172–190. doi:10.1016/j.cognition.2016.08.015

Rest, J. R., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999a). *Postconventional moral thinking: A Neo-Kohlbergian approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Rest, J. R., Narvaez, D., Bebeau, M., & Thoma, S. (1999b). A Neo-Kohlbergian approach: The DIT and schema theory. *Educational Psychology Review*, *11*(4), 291–324. doi:10.1023/a:1022053215271

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology*, *91*(4), 644–659. doi:10.1037/0022-0663.91.4.644

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi:10.3758/s13423-014-0595-4

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Thoma, S. J. (1986). Estimating gender differences in the comprehension and preference of moral issues. *Developmental Review*, *6*(2), 165–180. doi:10.1016/0273-2297(86)90010-9

Thoma, S. J. (2006). Research on the Defining Issues Test. In M. Killen & J. G. Smetana (Eds.),

Handbook of Moral Development (pp. 67–91). Mahwah, NJ: Psychology Press.

Thoma, S. J., Bebeau, M. J., Dong, Y., Liu, W., & Jiang, H. (2011). *Guide for DIT-2*.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1343-3

Walker, L. J., Gustafson, P., & Frimer, J. A. (2007). The application of Bayesian analysis to issues in developmental research. *International Journal of Behavioral Development*, *31*(4), 366–373. doi:10.1177/0165025407077763

Walker, L. J., Gustafson, P., & Hennig, K. H. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology*, *37*(2), 187–197. doi:10.1037//0012-1649.37.2.187

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. doi:10.3102/0034654311405999

Tables

Table 1

$P(H_0|D)$ for Jeffreys-Type prior from Berger & Sellke (1987).

		n						
		1	5	10	20	50	100	1,000
P-value	*t* statistic							
.10	1.645	.42	.44	.47	.56	.65	.72	.89
.05	1.960	.35	.33	.37	.42	.52	.60	.82
.01	2.576	.21	.13	.14	.16	.22	.27	.53
.001	3.291	.086	.026	.024	.026	.034	.045	.124

Table 2

Bayesian model comparison with the moral educational intervention dataset

Models	P(M)	P(M|data)	BF_M	$2logBF_M$	BF_{10}	$2logBF_{10}$	error %
Null model	.25	.013	.041	-6.39	1.000	.00	
With group assignment	.25	.062	.197	-3.25	4.610	3.06	5.701e -6
With pre-test engagement	.25	.097	.321	-2.27	7.243	3.96	.001
With both factors	.25	.828	14.473	5.34	62.009	8.25	.685

Note. P(M): prior model probability. P(M|data): posterior model probability. BF_M: BF comparing that model against all the other models. BF_{10}: BF comparing that model against the null model. error %: Size of numerical error associated with the Bayes factor.

Table 3

Analysis of Effects – Change in voluntary service activity in the moral educational intervention dataset

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$	$2logBF_{Inclusion}$
Group	.500	.890	4.599	3.05
Pre-test engagement	.500	.925	7.511	4.03

Note. P(incl): prior factor inclusion probability. P(incl|data): posterior factor inclusion probability. $BF_{Inclusion}$: BF of including a specific factor instead of not including the factor.

Table 5

Bayesian model comparison with the DIT-2 dataset

Models	P(M)	P(M|data)	BF_M	$2logBF_M$	BF_{10}	$2logBF_{10}$	error %
Null model	0.20	1.90e-94	7.61e-94	-428.83	1.00	.00	
With sex	0.20	2.04e-28	8.14e-28	-124.75	1.07e+66	304.08	3.53e-69
With grade level	0.20	2.56e-64	1.03e-63	-290.08	1.35e+30	138.75	.008
With both main effects	0.20	1.00	2725.08	15.82	5.25e+93	431.60	2.004
With the interaction effect	0.20	.001	.006	-10.28	7.71e+90	418.55	1.742

Note. P(M): prior model probability. P(M|data): posterior model probability. BF_M: BF comparing that model against all the other models. BF_{10}: BF comparing that model against the null model. error %: Size of numerical error associated with the Bayes factor.

Table 6

Analysis of Effects – DIT-2 dataset N2-score

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$	$2logBF_{Inclusion}$
Sex	.40	1.00	3.90e+63	292.85
Grade level	.40	1.00	4.91e+27	127.52
Interaction effect	.20	.001	.001	-13.05

Note. P(incl): prior factor inclusion probability. P(incl|data): posterior factor inclusion probability. $BF_{Inclusion}$: BF of including a specific factor instead of not including the factor.

Table 7

Results from classical and Bayesian correlation analyses

		P-score	N2-score	Age
N2-score	Pearson's *r*	.900		
	P-value	.001***		
	BF_{10}	∞		
	$2logBF_{10}$	∞		
Age	Pearson's *r*	-.029	-.027	
	P-value	.001***	.001***	
	BF_{10}	3,700.000	1,065.000	
	$2logBF_{10}$	16.432	13.941	
Grade level	Pearson's *r*	.049	.069	.443
	P-value	.001***	.001***	.001***
	BF_{10}	2.204e+14	7.711e+30	∞
	$2logBF_{10}$	66.053	142.240	∞

Note. BF_{10}: BF of presence of significant correlation. *** $p < .001$.

Figure Captions

Figure 1. Performing Bayesian t-test and ANCOVA with JASP. Users can simply select dependent and independent variables from the option panel (left). Once dependent and independent variables are set, JASP automatically reports outputs on the output screen (right). JASP's user interface is identical to that of SPSS.

Figure 2. Performing Bayesian correlation analysis with JASP.

Figure 3. Prior and posterior distribution of the Bayesian t-test of the moral educational intervention dataset.

---

[i] These were items found by using a keyword "Bayesian" from the JME: Derryberry and Thoma (2005), Heng, Blau, Fulmer, Bi, and Pereira (2017), Lee, Padilla-Walker, and Nelson (2015), and McGrath and Walker (2016).

[ii] This article demonstrates the result of Bayesian path analysis examining how motivational climate, basic psychological need, and moral disengagement influence antisocial and prosocial behavior in sport (Hodge & Gucciardi, 2015).

[iii] First, Kondo (1990) simulated the Bayesian Prisoner Dilemma game to model rational behavior, normative behavior, moral behavior, and cooperation. Second, Walker, Gustafson, and Hennig (2001) analyzed the consolidation and transition model in the development of moral reasoning measured by the Moral Judgment Interview with Bayesian techniques. Third, Walker, Gustafson, and Frimer (2007) overviewed benefits of Bayesian analysis and how to apply it in developmental psychology to address several methodological issues. Fourth, Railton's article (2017) discussed how moral learning occurs based on Bayesian perspective.

[iv] This tool can be downloaded for free from https://jasp-stats.org/.